www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20233-20236
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.12

A Study on Prediction of User Behavior Based on Web Server Log Files in Web Usage Mining

Anurag kumar[1], Vaishali Ahirwar[2], Ravi Kumar Singh[3]
[1, 2] Dr. APJ Abdul Kalam UIT,
Jhabua, MP, India
[1]*Anurag.davv@gmail.com*
[2]*vaishali07ahirwar@gmail.com*
[3]Prestige institute of Engineering Management and Research,
Indore, MP, India
Ravi.singh1308@gmail.com

Abstract: Nowadays, the growth of World Wide Web has exceeded a lot with more expectations. The internet is growing day by day, so online users are also rising. The interesting information for knowledge of extracting from such huge data demands for new logic and the new method. Every user spends their most of the time on the internet and their behavior is different from one and another. Web usage mining is the category of web mining that helps in automatically discovering user access pattern. Web usage mining is leading research area in Web Mining concerned about the web user's behavior. In this paper emphasizes is given on the user Behaviors using web server log file prediction using web server log record, click streams record and user information. Users using web pages, frequently visited hyperlinks, frequently accessed web pages, links are stored in web server log files. A Web log along with the individuality of the user captures their browsing behavior on a website and discussing regarding the behavior from analysis of different algorithms and different methods

Keywords: web usage mining, pre-processing, Apriori algorithm, FP-Growth algorithm, log files.

1. Introduction

The World Wide Web (WWW) is a huge resource of multiple types of information in varied formats which is very useful for the analysis of business progress, which is very important now days to stand in the competition of business. Researchers are beginning to investigate human behavior in this distributed Web data warehouse and are trying to build models for understanding human behavior in virtual environments. Data mining, often called Web mining when applied to the Internet, is a process of extracting hidden predictive information and discovering meaningful patterns, profiles, and trends from large databases. Web mining is an iterative process of discovering knowledge and is proving to be a valuable strategy for understanding consumer and business activity on the Web. There are three sub categories for mining web information. These sub categories are

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Web Content mining [3] deals with discovery of useful information from unstructured, semi structured or structured contents of web documents. Text, images, audio, video comprised by unstructured document, semi structured data includes HTML documents and lists and tables represent structured documents. The main aim of web content mining is to act as tool to retrieve information easily and quickly. Web Content Mining works by organizing a group of documents into related categories which helps web search engine to ex-tract information more quickly and efficiently. Web Structure Mining [6], [7] mines the information by utilizing the link structure of the web documents. It works on inter document level and discovers hyperlink structure. It helps in describing the similarities and relationships between sites. Web Usage Mining [3] is a data mining technique that mines the information by analyzing the log files that contains the user access patterns. Web Usage Mining mines the secondary data which is present in log files and derived from the interactions of the users with the web. Web usage Mining techniques are applied on the data present in web server logs, browser logs, cookies, user profiles, bookmarks, mouse clicks etc. This information is often gathered automatically access web log through the Web server.

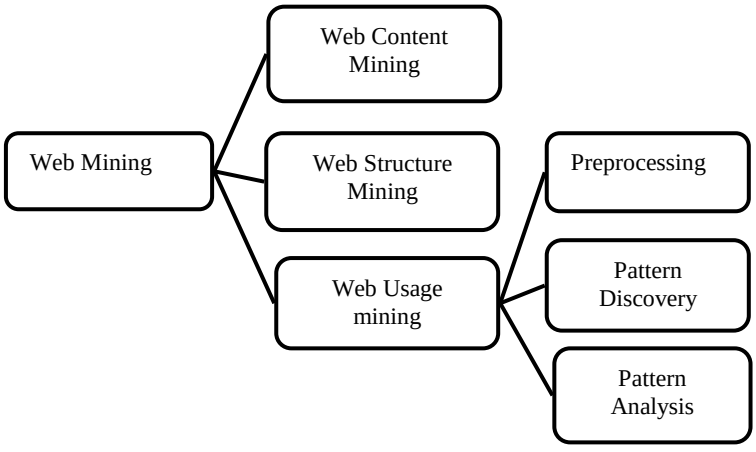

Fig. 1: Web Data Mining Structure

1.1 web usage mining

Web Usage Mining concentrates on the techniques that could predict the navigational pattern of the user while the user interacts with the web. It is mainly divided into two categories, they are general access pattern tracking and customized usage tracking. In general access pattern tracking information is discovered by using the history of web page visited by user while in customized usage

tracking mining is targeted on specific user. Mainly there are four types of data sources present in which usage data is recorded at different levels they are: client level collection, browser level collection, server level collection and proxy level collection.

Client Level collection: At this level data is gathered together by means of java scripts or java applets. This data shows the behavior of a single user on single site. Client side data collection requires user participation for enabling java scripts or java applets. The advantage of data collection at client side is that it can capture all clicks including pressing of back or reload button [2].

Browser Level Collection: Second method of data collection is by modifying the browser. It shows the behavior of single user over multiple sites. The data collection capabilities are enhanced by modifying the source code of existing browser. They provide much more versatile data as they consider the behavior of single user on multiple sites [2].

Server Level Collection: Web server log [5] stores the behavior of multiple users over single site. These log files can be stored in common log format or extended log format. Server logs are not able to store cached page views. Another technique used for usage data collection at server level is TCP/IP packet sniffing. Packet sniffers works by monitoring the net-work traffic and retrieve usage data directly.

Proxy Level Collection: Proxy servers are used by internet service provider to provide World Wide Web access to customers. These server stores the behavior of multiple user at multiple site. These server functions like cache server and they are able to produce cached page views. By predicting the usage pattern of the visitor Web Usage Mining improves the quality of e- commerce services, personalizes the web [1] or enhances the performance of web structure and web server.

Server data are data that are collected from web servers; it includes log files, cookies and explicit user input.
Servers contain different types of logs, which are considered to be the main date resource for web usage mining. The most popular logs are:

- Common Log Format (CLF): created to keep track of requests that occur on a website in chronological order. It contains the IP address of the client, hostname, username, time stamp, file name and file size. CLF has the following elements:
- Remote host: the IP address or domain name of the client
- Base URL: the URL of the user request
- Date: the date and time of the request
- Method: the method used by the client, such as GET, HEAD or POST
- File: the file requested by the client
- Protocol: the protocol used
- Code: the status code of the three requests; it consists of 3 digits
- Bytes: the number of bytes returned to the client
- Referrer: the URL from the referring server
- User agent: the operating system type and version

1.2 Why web usage mining?

The primary goal of web usage mining is to help people make good decisions to improve company performance and to maintain competitive advantage in the marketplace, i.e., it helps companies to make the best decisions quickly and easily. Web usage mining is the appropriate technique for extracting information and building a useful and knowledgeable database about customer behaviors. Also, it is very important in determining effective marketing strategies, i.e., those that increase sales and place the company's products on a higher level.

Therefore, it is easily determined that usage mining has valuable uses to predict web user behaviors. Each and every user thinking of thoughts and behavior is very different from one and another. This paper focused the web usage mining process of usage mining and algorithms used for usage mining and applying some data mining techniques.

2. Related work

The focus of related work to study and contrast the available technique to predict the web user behavior.

Jagan and Rajagopalan [9] describe the web usage mining and algorithms used for providing personalization on the web. In this paper focused the data preprocessing and pattern analysis on the web and using the association rule mining algorithms.

Ladekar A. Pawar A. *et al.* [10] describe a web mining algorithm that aims at amending the interpretations of the draft's output of association rule mining. This algorithm is being tremendously used in web mining. The results obtained prove the robustness of the algorithm proposed in this paper.

Parvatikar S. and Joshi B. [11] this paper focused on Web Usage Mining is the user navigation patterns and their use of web resources. The different stages involved in this mining process and with the comparative analysis between the pattern discovery algorithms Apriori and FP-growth algorithm.

Deepa and Raajan [12] implemented the preprocessing techniques to convert the log file into user sessions which are suitable for mining and reduce the size of the session file by filtering the least requested pages using the preprocessing technique. Data Preprocessing is one of the important tasks before applying mining algorithms. It converts the raw log file into user session. In this work, we have briefly introduced log file preprocessing and implemented it in a CTI log file. Also, we produce the summary of the user session file. We have used filtering technique to remove least requested resources.

Anand N. [13] describes an internet usage details and provides them with the tools to understand the online behavior of their teenage children. Singh A.P. and Jain R.C. [14] Different kinds of web usage mining techniques with their basic models and concepts are provided. In addition to that, for discovering the hidden patterns from web access log files a new model based on visual clustering is also suggested. The analysis of different methods of web usage mining.

Mishra R. and Choubey R. [15] describe the FPgrowth algorithm is obtaining a most frequently access paths and pages from the web log data and providing valuable information to user behavior.

Zubi Z. S. and Riani M.S.E. [16] discusses the use of web mining techniques is used to classify the web page's type according to user visits. This classification is helps to understand the web user behavior. The classification and association rule techniques for discovering the interesting information from browsing patterns.

Avneet Saluja *et al.* [17] in their work is user future request prediction using web log records and user information. The purpose of the effort is to provide a benchmark for evaluating a various methods used in the past, a present and which can be used in a future to minimize the search time of a user on the network.

3. Proposed work

The proposed system aims to discover the web user behavior from users through the web server log files. In the prediction model, the client requests the web page and links automatically get stored in a server log file. All requested web page links get stored and maintained by the client. A more noisy and dirty data's present in the log files. Preprocessing steps are necessary to remove the noise, missing and redundant data. During preprocessing, the client retrieves frequent web link used by users, the HTTP request pages, user identification in different users and sessions is identified and pattern completion. Using frequent web links, we predict the user behavior and identify what are all the sites mostly viewed by users. These techniques used to predict the user's behavior.

4. Methodology

4.1 Association rule Mining

A various prediction methods are available to predict the web user behavior. Mining of association rules is imperative research in web usage mining. There are many algorithms are applied in web usage mining. Association rule mining is an innovation and correlation a set of frequent item sets or pages. In most frequent algorithms are Apriori, Apriori TID, STEM, DIC, Partition Algorithm, Eclat and FPGrowth, etc. Since the algorithms are mining frequent item sets.

a. Apriori algorithm

Apriori algorithm is a step-wise search; the n number of item sets is used to discover an n+1 item sets. A set of usual sets, scanning the database to gather the count for each item, and collecting those item sets is satisfying the minimum support. The resulting set is denoted as one common item set.

Next, the common time set is used to find next interesting item sets, this process is continue until get the most frequent item sets. A final iteration, you will end up with many n-item sets, this is basically called association rules. To pick interesting rules from the set of all prospective rules and various constraint measures such support and confidence is applied.

b. FP-Growth

This algorithm [15] calculates frequent items sets from large data sets. The advantage of Frequent Pattern Growth is compared to Apriori is that it uses only two times data scans is therefore often valid even on large data sets.

c. Frequent pattern discovery

Pattern Discovery is one of the Web Usage Mining process. It is one of the techniques in association rule mining, According to the most recent of web log analysis methodology; data mining methods are more reliable and efficient for the hidden pattern discovery. But most of the authors are working on supervised methods. Unsupervised methods are not finding the pattern extraction from web logs.

D. Pattern Analysis

Pattern Analysis is the goal of the process. In this process is used to extract the appealing patterns from the log files. It is one of the pattern discovery process. Reject a un-relevant data from web log files. Pattern discovery and pattern analysis are the preprocessing stages in web usage mining.

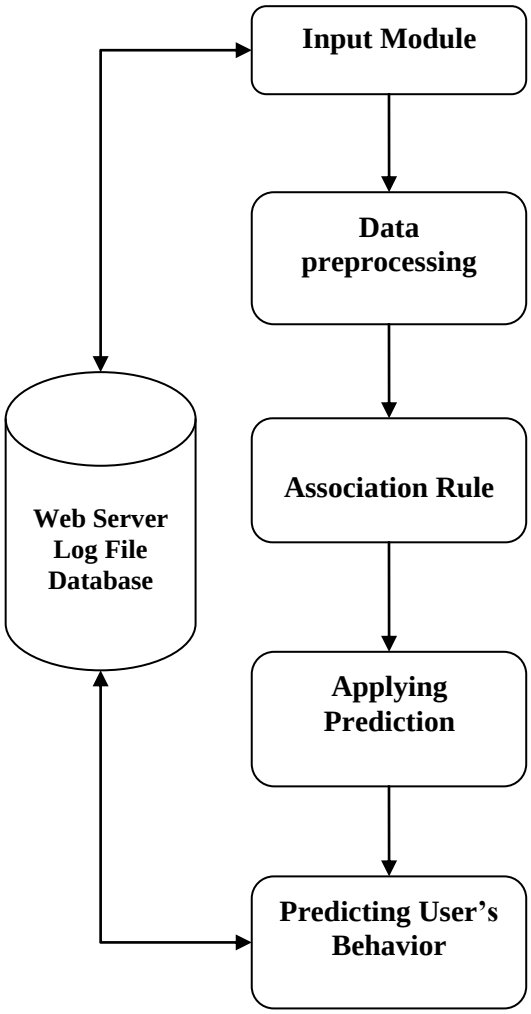

Fig. 2: Architecture of Proposed System

5. Conclusion

From our study various research papers on basis web usage mining we conclude that web usage mining is extracting the information from web server log file which access by users.

Data Mining is the process to mine the interesting knowledge from the enormous amount of data. In previous work, they predict the future request given by the user based on the current request, analyze the pattern and they also predict the kids' behavior. In this paper, we discuss various methods of data mining techniques and describe the way how to apply data mining techniques. The proposed work is to develop a prediction model to predict user behavior based on web server log files. The client analyzes the links used by users and predicts the behavior of the user based on the websites they are used.

References

[1] G. Salton and M.J. McGil.,"Introduction to modern Information Retrieval.", McGraw-Hill, New York, 1983.

[2] J. Srivastava, R. Cooley, M. Deshpande and P. Tan., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web data", Department of Computer Science and Engineering, University of Minnesota. SIGKDD Explorations, 1(2):12, January 1999.

[3] Kosla, R. and Blockeel, H.2000, "Web Mining Research: A Survey", SIGKDD Explorations Vol. 2, 1-15.

[4] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," in Proc. Ninth IEEE International Conference on Tools with Artificial Intelligence, 1997, pp. 558-567.

[5] R.Shanthi, Dr.S.P.Rajagopalan, "An Efficient Web Mining Algorithm To Mine Web Log Information", IJIRCCE Vol. 1, Issue 7, September 2013.

[6] H. Han and R. Elmasri.," Learning Rules for Conceptual Structure on the Web", J.Intell. Inf. Syst. Vol.22, No3 pp 237-256, 2004.

[7] J. Hou and Y. Zhang," Effectively finding relevant Web Pages from Linkage Information", IEEE Trans. Know. Data Eng. Vol. 15, No.4, pp-940-951, 2003

[8] G. Salton and M.J. McGil.,"Introduction to modern Information Retrieval.", McGraw-Hill, New York, 1983.

[9] S. Jagan, and S.P. Rajagopalan, "A survey on web personalization of web usage mining", IRJET International Research Journal of Engineering and Technology, 2015.

[10] A. Ladekar, P. Pawar, D. Raikar and J. Chaudhari, "Web Log Based Analysis of User's Browsing Behavior", IJCSIT - International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015.

[11] S. Parvatikar and B. Joshi, "Analysis of User Behavior through Web Usage Mining", ICAST – International Conference on Advances in Science and Technology, 2014.

[12] A. Deepa, and P. Raajan, "An efficient preprocessing methodology of log file for Web usage mining", NCRIIAMI - National Conference on Research Issues in Image Analysis and Mining Intelligence, 2015.

[13] N. Anand, "Effective prediction of kid's behavior based on internet use", International Journal of Information and Computation Technology, 2014.

[14] A.P. Singh, R. C. Jain, "A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation", IJETTCS - International Journal of Emerging Trends & Technology in Computer Science, Vol 3, 2014.

[15] R. Mishra, A. Choubey, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 2, 2012.

[16] Z.S. Zubi, M.S. Riani, "Applying web mining application for user behavior understanding", Recent Advances in Image, Audio and Signal Processing.

[17] Saluja, B. Gour, and L. Singh., "Web Usage Mining Approaches for User's Request Prediction: A Survey" IJCSIT-International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015

[18] H.N. Randhir, R. Gupta, "WUM for Browsing Behavior of a User and Subsequently to Predict Desired Pages: A Survey", IJESIT -International Journal of Engineering Science and Innovative Technology, Vol 2, 2013.

[19] S. Khan, Y. Singh and A. K. Sachan., "Web Mining Approach in Analysing User Behaviour and Interest for Website Modification", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 5, 2015.

[20] A. Vishwakarma, and K.N. Singh, "A survey on web log mining pattern discovery", IJCSIT – International Journal of Computer Science and Information Technologies, pp: 7022-7031, 2014.

[21] S.P. Ajeetkumar, P.K Anagha, "Review on Exploring User's Surfing Behavior for Recommended Based System", IJETTCS - International Journal of Emerging Trends & Technology in Computer Science, Vol , 2014.